A Statistical Monitoring and Diagnosis System for High-Throughput DNA Sequencing

Mingkun Li^{1}, Alex C. Copeland^{1}, Susan Lucas^{1}

Keywords: Process monitoring and diagnosis, DNA sequencing, Information fusion, Visualization

1 Introduction.

One key constraint for large scale, high-throughput DNA sequencing is cost. While there are many promising emerging sequencing technologies, Sanger sequencing and capillary electrophoresis remain the work horses of major sequencing centers. The sequencing process involves many steps with different templates, reagents, instruments, and operators and it is not stable. To minimize the cost and increase throughput, it is necessary to quickly find failures and their causes when they occur. Given the complexity of this process, it is not uncommon that human operators spend hours even days to find and diagnose failures.

This poster presents a statistical monitoring and diagnosis system for high-throughput DNA sequencing. Statistical techniques are used to model the sequencing process. Various information is integrated to make a thorough decision. Visualization is used to help user better understand the process. This system is currently being used in the DOE Joint Genome Institute, and it reduces the time required for detecting and diagnosing problems from hours to minutes.

2 Statistical modeling of sequencing process.

In our process, DNA sequencing is performed in a plastic plate which has 384 wells, where each well contains a unique clone. All clones in a plate nominally experience identical processing. A primary quality metric of sequencing is the high quality read length (Q20 read length) for each plate or clone, which measures the total number of bases having error probabilities less than 1 in 100. There are two main tasks in monitoring and diagnosis. The first is to classify whether a sequenced plate is a failure or not, and the second is to classify whether a capillary is broken. Then, failed plates or wells can be aggregated to find the source of the failures using other information.

For simplicity, it is assumed that for clones coming from the same library, each plate's Q20 read length follows a Gaussian distribution if there are no sequencing failures. Therefore, plates with low Q20 read length are most likely caused by sequencing process errors, e.g. machine malfunction. If the parameters (mean and standard deviation) of the Gaussian distribution are known, a plate can be classified as a failure or not with specified confidence. The sample mean and standard deviation are usually used as the estimators for the mean and standard deviation. However, the distribution of Q20 read length for a plate is typically shaped like figure 1, a small tail near 0 and a dominant bell in the middle. It is reasonable to assume that the lower performing plates (near tail) are caused by production failure and the dominant bell shape is the inherent property of the library. Thus it is more accurate to estimate mean and standard deviation using only good plates. This can be achieved by recursively using a Gaussian distribution. First, the sample mean and variance are calculated; then any samples with Q20 read length one standard deviation lower than

[1] DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA
E-mail: mli@lbl.gov

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

LBNL-59411 Abs.
UCRL-

the mean are removed, and the remaining samples are used to calculate sample mean and variance, which are the final estimators for mean and variance.

The Q20 read length of a clone is also assumed to follow a Gaussian distribution. Any well is classified as bad if this well has a Q20 read length 3 standard deviation less than the sample mean. The probability that a sample has a Q20 read length 3 standard deviation less than the mean is 0.0014, and the probability that one well has a Q20 3 standard deviation less than the mean for a 384 plate is 0.52. Thus, any well with a Q20 read length 3 standard deviation less than the mean is very likely caused by sequencing failure such as clogged tips and broken capillaries. Another clue is that there are four wells for each capillary on the ABI sequencing machines we use. Thus, when the 3 or 4 wells corresponding to a capillary are under performing, it is almost certain that the capillary is broken (the red in Figure 2).

3 Reports.

The system generates a sequencing report everyday for operators. The report consists of the three parts. First, it summarizes the sequencing information of large libraries, and reports organism, vector, number of runs, average Q20 read length, average fail rate, and average signal intensity. Next it presents the information on low performance sequencers, such as number of runs, number of bad runs, average Q20 read length, average fail rate, number of broken capillaries, number of runs for the current array, and average signal intensity. The overall array pattern is also presented to aid in quickly finding array patterns, if there any. The relevant detail information is linked to summaries, making it easy to drill down if desired. Sequencing data is also available for download to perform user specific analysis.

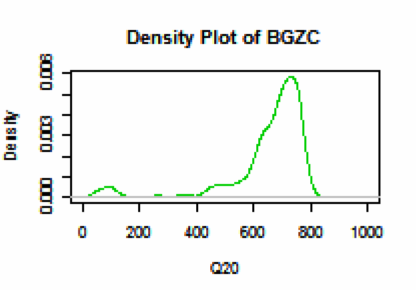
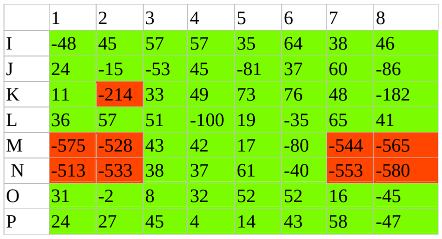


Figure 1. Density plot of the library BGZC



	1	2	3	4	5	6	7	8
I	-48	45	57	57	35	64	38	46
J	24	-15	-53	45	-81	37	60	-86
K	11	-214	33	49	73	76	48	-182
L	36	57	51	-100	19	-35	65	41
M	-575	-528	43	42	17	-80	-544	-565
N	-513	-533	38	37	61	-40	-553	-580
O	31	-2	8	32	52	52	16	-45
P	24	27	45	4	14	43	58	-47

Figure 2. Part of the well difference map of a sequencer

References

[1] Ewing, B., Green, P. 1998. Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* 8:186-

[2] Ewing, B., Hillier, L., Wendl, M., Green, P. 1998. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* 8:175-185

[3] Montgomery, D. C. 1996. *Introduction to Statistical Quality Control, 3rd edition*. John Wiley & Sons

[4] Montgomery, D. C. 2000. Design and Analysis of Experiment, 5th edition. John Wiley & Sons

[5] Li, M., Feng, S., Sethi, I. K., Luciow, J., Wagner, K. 2003. Mining production data with neural network & CART. In: Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03). Melbourne, Florida

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.

LBNL-59411 Abs.
UCRL-